From Evidence to Belief: A Bayesian Epistemology Approach to Language Models

Anonymous ACL submission

Abstract

This paper investigates the knowledge of language models from the perspective of Bayesian epistemology. Specifically, it aims to explore whether language models can accurately incorporate evidence of varying levels of informativeness and reliability into their confidence and responses. As Bayesian epistemology interprets belief as confidence according to evidence, this study offers a new perspective on understanding the beliefs and knowledge of language models. We created a dataset with various types of evidence and analyzed its response and confidence using verbalized confidence, token probability, and sampling. From the perspective of verbalized confidence, our research has shown that we can interpret that language models can generally reflect evidence in their confidence and calibration. We also demonstrated that language models exhibit biases toward correct evidence, exploit unreasonable evidence, and ignore errors in the context, all of which can be interpreted as the epistemic character of language models.

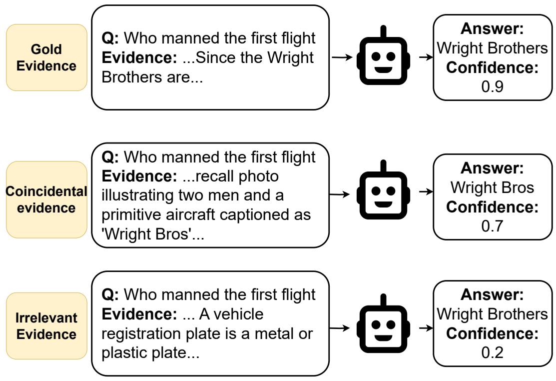

Figure 1: We explored changes in the confidence and responses of language models by providing them with various types of evidence. For evidence, we used verbalized confidence (Tian et al., 2023b), token probability and sampling method.

1 Introduction

Large Language models (LLMs) have advanced to the point where they can naturally respond to various practical tasks such as question-answering and conversation (OpenAI et al., 2023; Gemini Team et al., 2024). However, limitations like hallucination and trustworthiness still exist, and research efforts continue to address these issues (Huang et al., 2023; Sun et al., 2024; Xiao and Wang, 2021; Zhang et al., 2023). In this paper, we take a different approach by examining language models from a philosophical motivation. "Do language models possess knowledge?" in other sophisticated words, "Can we interpret language models as possessing knowledge?" Knowledge is generally defined as justified true belief. "s knows p" means that (1) p is true, (2) p is justified by s, and (3) p is a belief (Audi, 1997). Most AI research has focused on aspect (1), that is, whether the language model's response is true or not, using metrics for measuring correctness such as accuracy (Thorne et al., 2018; Hendrycks et al., 2021; Srivastava et al., 2023). Explanation generation (Wei et al., 2023; Camburu et al., 2018) can be interpreted as (2), exploring the language model's ability to provide justification. This paper addresses (3), the belief of language models. Specifically, it deals with the relationship between belief and the language model's justification, expressed as evidence. Since belief is a vague and challenging concept to define, this paper focuses on belief from the perspective of Bayesian epistemology, which interprets belief as a quantitative and functional variable.

According to Bayesian epistemology, the degree of belief can be interpreted and measured as probability, called *probability norm*. In particular, regarding the confirmation of belief, we should adjust the confidence of belief based on the evidence. Specifically, when H represents the hypothesis, which can be interpreted as belief, E means the evidence for the belief, and θ represents the background information or prior knowledge, we can define 3 assumptions:

- **Confirmation Assumption:** E confirms H if and only if $P(H \mid E, \theta) > P(H \mid \theta)$
- **Disconfirmation Assumption:** E disconfirms H if and only if $P(H \mid \theta) > P(H \mid E, \theta)$.
- **Irrelevance Assumption:** E is irrelevant to H if and only if $P(H \mid \theta) = P(H \mid E, \theta)$.

Also, if we define belief in terms of probability, the strength of the evidence should also be reflected in the confidence. That is,

- **Evidence Power Assumption:** E' confirms H more stronlgly than E'' if and only if $P(H \mid E', \theta) - P(H \mid \theta) > P(H \mid E'', \theta) - P(H \mid \theta)$.

which is equivalent with $P(H \mid E', \theta) > P(H \mid E'', \theta)$ (Horwich, 1982; Howson, 2000; Talbott, 2006; Hájek and Hartmann, 2010).

The degrees of belief should not only be a probability. The probabilities assigned to these beliefs must align with the *calibration norm*, meaning they should correspond to the actual likelihood of the event occurring, that is, the actual frequency (Williamson, 2010).

The goal of this paper is to explore whether various types of evidence are reflected in language models' confidence and responses. The evidence here is not merely perturbations altering the correctness of information, i.e., informativeness, but also includes a dataset of various types of evidence modified for reliability factors such as coincidence, timeliness, source of credibility, etc.

We observed that language models generally form justified beliefs that align with Bayesian assumptions. However, we also identified epistemic traits of language models, such as a bias towards golden evidence and tendencies to utilize unreasonable information, ignore inaccuracies, or be hindered by excessive specificity.

2 Related Works

Calibration of LLMs Calibration of language models has long been considered an important metric for faithful AI, with the log probability of neural models being regarded as the confidence in their responses (Kadavath et al., 2022; Guo et al., 2017). As language models have grown in size, research has also emerged on verbalized confidence, where the models themselves generate confidence in their responses (Lin et al., 2022; Mielke et al., 2022; Tian et al., 2023b). While confidence can be used to enhance the performance of language models (Zhao et al., 2023; Tian et al., 2023a), there is also research focused on the interpretation of the models' confidence itself. For example, Kuhn et al. (2023) measured model uncertainty using semantic space, and Xiong et al. (2024) defined confidence through various prompts and sampling methods.

The study most similar to ours is Zhou et al. (2023), which investigated the impact of epistemic markers on model calibration. However, unlike their focus on linguistic markers, our paper examines how changes in epistemic evidence, containing information on both content and reliability, influence confidence and calibration.

Belief and epistemology of LLMs Research on the belief of LLMs has primarily focused on whether these models maintain consistent beliefs. Kassner et al. (2023) constructed belief graphs for LLMs and examined whether using these belief graphs improved responses. Hase et al. (2023) experimented with input paraphrase, entailment methods, and belief graph construction to determine if models possess beliefs. Kassner et al. (2021) argued for the necessity of storing consistent information for the beliefs of LLMs. van Dijk et al. (2023) interpreted LLMs from a philosophical pragmatism viewpoint, while Kim and Thorne (2024) suggested that LLMs might not be epistemologically holistic by showing that they fail to preserve core knowledge effectively. This paper also addresses the epistemological aspects of LLMs, specifically concerning belief. However, it aims to measure not only the content of belief but also its degree.

Adversarial Context With the advent of in-context learning, many studies have investigated the impact of few-shot demonstrations and explanations on generated responses (Brown et al., 2020; Wei et al., 2022). Wang et al. (2023a) indicated that even inaccurate demonstrations could be utilized in Chain-of-Thought (COT) prompting. Chia et al. (2023) improved question accuracy through contrastive demonstrations, and Chen et al. (2023) explored the effect of the number of demonstrations on accuracy. While these papers discuss the impact of demonstrations on accuracy, we aim to explore how the direct evidence of question influences not only the accuracy but also the confidence and calibration of language models.

Turpin et al. (2023); Lanham et al. (2023) experimented with various perturbations in generated COT inputs and their effects on answers, which is similar to our approach. While these studies modified explanations based on informativeness (such as incorrectness or relevance), our paper aims to investigate whether LLMs can reflect various evidence on their confidence and calibration. In addition, we have explored how epistemically diverse evidence, such as coincidental evidence and evidence from sources of varying credibility, affects the model's confidence.

3 Methods

As Figure 1, our experiment provides various types of evidence as context to language models and then observes its confidence and responses. Influenced by Bayesian epistemology, we defined a confirmation task to measure whether language models can reflect the confirmation, disconfirmation, or irrelevance assumption introduced in section 1. Also, we created a strength of evidence task to assess LLM's ability to represent the various power of evidence. To measure the probability norm for adjusting confidence according to the evidence, we used an average confidence across all samples. In order to measure the response, such as correctness or calibration norm, we used accuracy (ACC) and Expected Calibration Error (ECE).

3.1 Experimental Design

We estimated the confidence of language models using verbalized confidence (Verb. 1S top-1) (Tian et al., 2023b), token probability, and sampling (Lee et al., 2023; Xiong et al., 2024) (See Appendix E.2 and F.1 for detail). Smaller-scale open-source LLMs struggled to generate responses in the correct format matching the prompt of verbalized confidence. Also, Tian et al. (2023b) mentioned that closed-source models are better at generating verbal confidence than open-source models. Therefore, we used GPT-3.5-turbo-0125 and GPT-4o-2024-05-13 for inference. We used SciQ (Welbl et al., 2017), TriviaQA (Joshi et al., 2017), GSM8K (Cobbe et al., 2021) for inference and making evidence dataset for Confirmation task, and used only SciQ dataset for Strength of Evidence task, as a scientific question is suitable for making various degree of reliable evidence (See Appendix E.3 for dataset statistics.).

3.2 Confirmation Task

The objective is to observe and analyze the changes in the language model's confidence and responses when presented with various types of evidence, compared to scenarios where the language models receive the original evidence E or in the absence of E, and assess how these changes align with three assumptions: Confirmation, Disconfirmation, and Irrelevance introduced section 1. Let the entire dataset be

$$\begin{aligned} D = \{S_i = (Q_i, A_i, E_i) \mid\ & Q_i \text{ is a question}, \\ & A_i \text{ is an answer for } Q_i, \\ & \text{and } E_i \text{ is evidence for } Q_i \text{ and } A_i\}. \end{aligned} \tag{1}$$

and

$$E_i = (s_{i1}, s_{i2}, \dots, s_{in}) \tag{2}$$

where s_{ij} is a sentence of E_i and $j = \{1, \dots, n\}$ (index of sentence in E_i). For the experiment, we need to create modified (Q_i, A_i, E_i') where E_i' is a perturbation of E_i. The following are the types of E_i':

1. **Negated Evidence**
   Evidence where sentences in E_i are replaced with their negated sentences. Thus, E_i' is negated evidence if and only if

   $$E_i' = (\neg s_{i1}, \neg s_{i2}, \dots, \neg s_{in}) \text{ for all } s_{ij} \in E_i.$$

2. **Incomplete Evidence**
   Evidence that includes only a subset of sentences from the original evidence E_i. Thus, E_i' is a proper subset of E_i. We used E_i', which contains only 50% of the sentences from the E_i in our main experiment.

3. **Contradictory Evidence**
   Original evidence E_i with additional negated sentences from E_i. Thus, E_i' is contradictory evidence if and only if

   $$E_i' = E_i \cup N \quad \text{where} \quad N \subset \{\neg s_{ij} \mid s_{ij} \in E_i\}$$

   such that $|N| = 0.5 \times |E_i|$. That is, adding 50% of the negated evidence to the original evidence.

4. **Irrelevant Evidence**
   Irrelevant evidence is $E_i' = E_j$ where $j \neq$

i. That is, E_i is randomly shuffled within the dataset D so that the evidence E_i of tuple (Q_i, A_i, E_i) is replaced with evidence E_j from a different tuple (Q_j, A_j, E_j).

5. **Coincidental Evidence**
For the SciQ and TriviaQA dataset, unlike other previous types of evidence, coincidental evidence does not include incorrect answers but explanations reaching the golden answer by irrational reasoning or epistemic luck. Examples include explanations derived from random guessing or vague memories. For GSM8K, coincidental evidence consists of a wrong reasoning process but a correct final answer.

3.3 Strength of Evidence

This task differs from the Confirmation task in that it focuses on the strength of evidence. Unlike the modified E' used in the Confirmation task, the evidence used here includes the correct answer but perturbation of reliability. The goal is to understand how differences in the strength of evidence impact confidence and calibration and assess whether LLMs align with Evidence Power Assumption in section 1. For each (Q_i, A_i) pair, two types of perturbation (Q_i, A_i, E_i') and (Q_i, A_i, E_i'') are created. E_i' represents more reliable evidence, while E_i'' represents relatively less reliable evidence. The following are the types of evidence:

1. **Source of Credibility**
For each (Q_i, A_i) pair, E_i' means evidence from a highly reputable and authoritative source, while E_i'' means evidence from an anonymous online post or an individual.

2. **Specificity and Detail**
This involves varying the detail and specificity of the evidence. Similar to source of credibility, for each (Q_i, A_i), E_i' is highly detailed evidence, while E_i'' is evidence with general mentions related to the question.

3. **Timeliness**
This involves modifying the evidence based on its recency. For each (Q_i, A_i), E_i' consists of recent findings and experiments, while E_i'' consists of relatively older findings and experiments.

4. **Experimental Evidence**
For each (Q_i, A_i), E_i' includes evidence derived from precise and controlled experiments, while E_i'' includes evidence where the answer is observed by a witness without experiments.

You can see the prompt for generating the evidence in Appendix F.2

4 Results and Analysis

4.1 LLMs on Confirmation task

You can see the results of the Confirmation task using the verbalized confidence method in Table 1. The results for the token probability method and the sampling method are presented in Table 2 and Table 3, respectively, both of which are located in Appendix A.

LLMs follow confirmation assumption In Table 1, 2 and 3, NO_EVI and EVI column show that when E is golden evidence that helps confirm the answer, we observe $P(H \mid E) > P(H)$ across all models, datasets and methods we used, which align well with the Confirmation assumption of Bayesian epistemology. Moreover, except for a slight increase in ECE when golden evidence is present in the case of GPT-3.5 on Trivia QA with verbalized in Table 1 and sampling method in Table 3, both ACC and ECE showed good results when given such confirming evidence. This indicates that language models have strong confidence and handle information well when the evidence contains purely helpful information for deriving the correct answer. We can interpret that language models satisfy the probability norm and calibration norm in the confirmation case. Excluding GSM8K, in NO_EVI colimn, we can see that the language model has some degree of parametric knowledge about SciQ and TriviaQA. However in GSM8K, the average confidence and accuracy significantly improve when evidence is provided, and ECE significantly decreases. This shows that language models cannot complex reason well without any explanation and reaffirms the importance of explanation in arithmetic tasks (Wei et al., 2023).

Case of disconfirmation: Negated evidence and Contradictory evidence In the verbalized method in Table 1, except for the GSM8K in no evidence baseline on GPT-3.5, which performs poorly on all metrics, negated evidence (Negation) shows low confidence, low accuracy, and high ECE compared to all no-evidence baselines, which is well-aligned with bayesian assumption on disconfirmation case. Low confidence indicates that LLMs do

	Dataset	Metric	No_EVI	EVI	Coincidence	Irrelevant	Negation	Incomplete	Contradiction
GPT-3.5-turbo	SciQ	Confidence	0.851	0.943	0.835	0.714	0.827	0.928	0.945
		Accuracy ↑	0.67	0.841	0.854	0.53	0.572	0.77	0.847
		ECE ↓	0.18	0.111	0.071	0.262	0.304	0.161	0.108
	Trivia	Confidence	0.827	0.922	0.818	0.69	0.797	0.897	0.925
		Accuracy ↑	0.846	0.879	0.971	0.698	0.702	0.86	0.869
		ECE ↓	0.035	0.058	0.153	0.125	0.211	0.06	0.076
	GSM8K	Confidence	0.74	0.998	0.988	0.765	0.931	0.96	0.949
		Accuracy ↑	0.078	0.951	0.843	0.066	0.023	0.666	0.777
		ECE ↓	0.662	0.048	0.148	0.699	0.911	0.307	0.197
GPT-4o	SciQ	Confidence	0.925	0.986	0.902	0.861	0.875	0.948	0.977
		Accuracy ↑	0.73	0.915	0.88	0.7	0.675	0.82	0.905
		ECE ↓	0.195	0.073	0.04	0.171	0.2	0.128	0.072
	Trivia	Confidence	0.915	0.933	0.895	0.878	0.866	0.909	0.926
		Accuracy ↑	0.94	0.96	0.99	0.935	0.86	0.945	0.955
		ECE ↓	0.037	0.027	0.095	0.063	0.048	0.036	0.037
	GSM8K	Confidence	0.924	0.991	0.83	0.89	0.883	0.96	0.957
		Accuracy ↑	0.24	0.97	0.54	0.195	0.165	0.774	0.96
		ECE ↓	0.684	0.033	0.406	0.705	0.718	0.186	0.013

Table 1: The result of confirmation task with verbal confidence methods. We used 200 samples for GPT-4o due to the cost limit. NO_EVI refers the question with no context which means $P(H \mid \theta)$, serving as baseline. Others are the case of $P(H \mid E, \theta)$ where evidence appears in the context. EVI refers to the context in which the golden evidence from the dataset is given, while the other evidence types are those mentioned in section 3.2.

not simply follow the negated evidence to generate an answer, but rather that the negated evidence conflicts and confuses with existing parametric knowledge, which leads to lower accuracy and higher ECE.

However, in the case of Token probability and Sampling method, when Negated Evidence is presented, the ACC decreases, and the ECE increases in most cases, but the Confidence inconsistently decreases or increases compared to the baseline. That is, in the disconfirm case, both sampling and token probability fail to reflect the degree of belief according to the evidence adequately.

On the other hand, in most models and methods, contradictory evidence, which contains both correct and negated evidence in the context, shows higher confidence and accuracy than the no-evidence baseline in all cases except for some results of the TriviaQA dataset, which shows slightly lower ACC and slightly higher ECE. Surprisingly, despite the presence of inaccurate information, the model appears high-confident and well-calibrated in almost all scenarios. This indicates that the language models can effectively filter the given context and generate responses without conflicting with its parametric knowledge. Unlike the case with negated evidence, it can be interpreted that the existence of incorrect sentences is offset by the influence of golden evidence. Hence, language models do not consider contradictory evidence as evidence for disconfirming the beliefs.

The verbalized method can reflect not only the unreliability but also the information of coincidental evidence. When language models receive coincidental evidence as input, except for the notably low performance of GPT-3.5 with no evidence on GSM8K, the average confidence of the verbalized method decreased compared to no evidence in verbalized method. This means that, although the evidence contains the correct answer, the LLMs understands that the method to reach that answer is unreliable and unreasonable through verbalized confidence, thus decreasing its degree of belief.

However, compared to the baseline with no evidence, we observe that accuracy increases when coincidental evidence is present. This shows that LLMs generate responses using correct answer in the evidence, regardless of its poor reliability and irrationality. Additionally, except for Trivia QA, ECE decreases when coincidental evidence is given compared to the no evidence baseline. This means that, although the LLMs shows slightly lower confidence in its responses, the responses generated using this evidence align well with the correct answers and have a higher frequency of being correct. We interpret that SciQ and GSM8K are more challenging than Trivia QA, and thus, the LLMs exhibit conservative confidence responses when faced with less reliable evidence for these datasets.

On the other hand, in the Token and Sampling method (Table 2, 3), confidence has increased across the board, which means this method fails

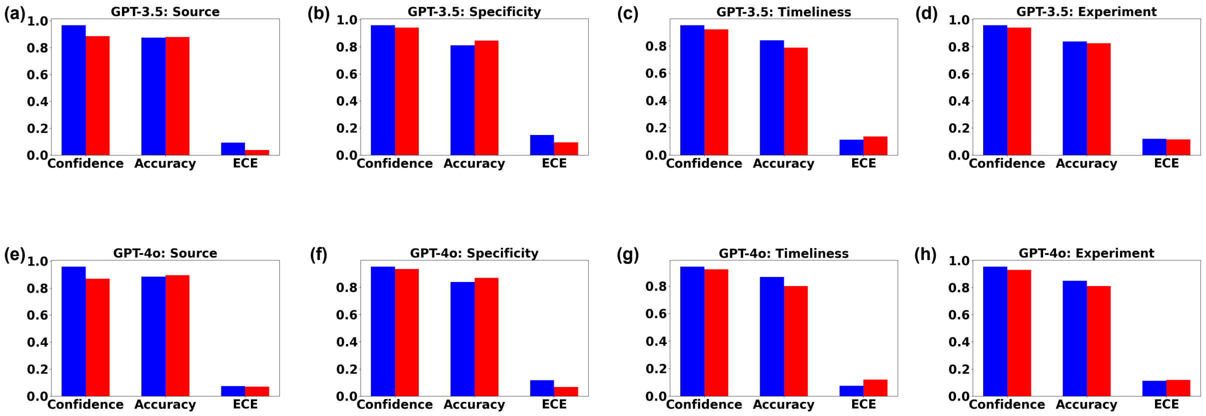

Figure 2: The results of the Strength of Evidence task on the SciQ dataset. The blue bar represents the cases where the strength of evidence is high. Specifically, the blue bar indicates the context from more credible sources, more specific, recent, and experimental evidence, while the red color represents less credible sources, less specific, old, and observational evidence. We found that, in all models and datasets, higher power of evidence leads to greater confidence with verbalized confidence. However, it does not always result in improvements in accuracy and ECE.

to capture the reliability of evidence. Additionally, ACC has increased and ECE has decreased. Thus, the Token and Sampling method fails to recognize the trait of coincidence and interprets it as typical confirmation evidence. We have interpreted verbalized confidence as a method in which, unlike other confidence methods, it explicitly requires the LLMs to generate confidence, reflecting various aspects of evidence in the confidence level. Thus, verbalized confidence acts as a form of introspection function.

Incomplete evidence acts as a positive hint. When incomplete evidence is provided, the confidence in the language model's response increases except for a slight decrease in GPT-4 on Trivia QA with the verbalized method. Except for some TriviaQA cases, accuracy also increases, and ECE decreases for all models and methods. Incomplete evidence does not contain inaccurate information and can be considered as a partial subset of the gold evidence, acting as a hint. Similar to the contradictory evidence case, we can see that the language model is biased towards imperfect golden evidence.Therefore, while not as effective as golden evidence, the language model reflects the information from the evidence well without distraction.

LLMs are highly confused by irrelevant evidence According to Bayesian epistemology, confidence should not change when irrelevant evidence is provided. However, even considering that this equation in irrelevant contexts might be too rigid for probabilistic language models, the results for verbalized method show that, except for GPT-3.5 on GSM8K, the average confidence and accuracy significantly decrease and ECE significantly increases when irrelevant evidence is provided compared to when no evidence is given, across most models and datasets (see Table 1).

Similarly, in the Token probability method, average Confidence and ACC have decreased in all cases, and excluding the GSM8K case, ECE has mostly increased. In the Sampling method as well, excluding some cases of TriviaQA and SciQ with GPT-4o, both confidence and ACC have decreased, and ECE has shown a tendency to increase.

This indicates that language models are severely distracted by irrelevant text in terms of the content of the evidence as in (Shi et al., 2023). Unlike contradictory evidence, the inability to filter out such irrelevant evidence leads to cognitive confusion, resulting in lower accuracy and reduced confidence.

4.2 LLMs on Strength of Evidence task

You can see the results of the Strength of Evidence task using the verbalized confidence method in Figure 2. The results for the token probability method and the sampling method are presented in Figure 4 and Figure 5, respectively, both of which are located in Appendix B.

High credible, highly detailed evidence can give confidence, but not accurate response in verbalized confidence As in (a), (b), (e), (f) in Figure 2, when the evidence comes from a credible source or includes more detailed explanations, we observe

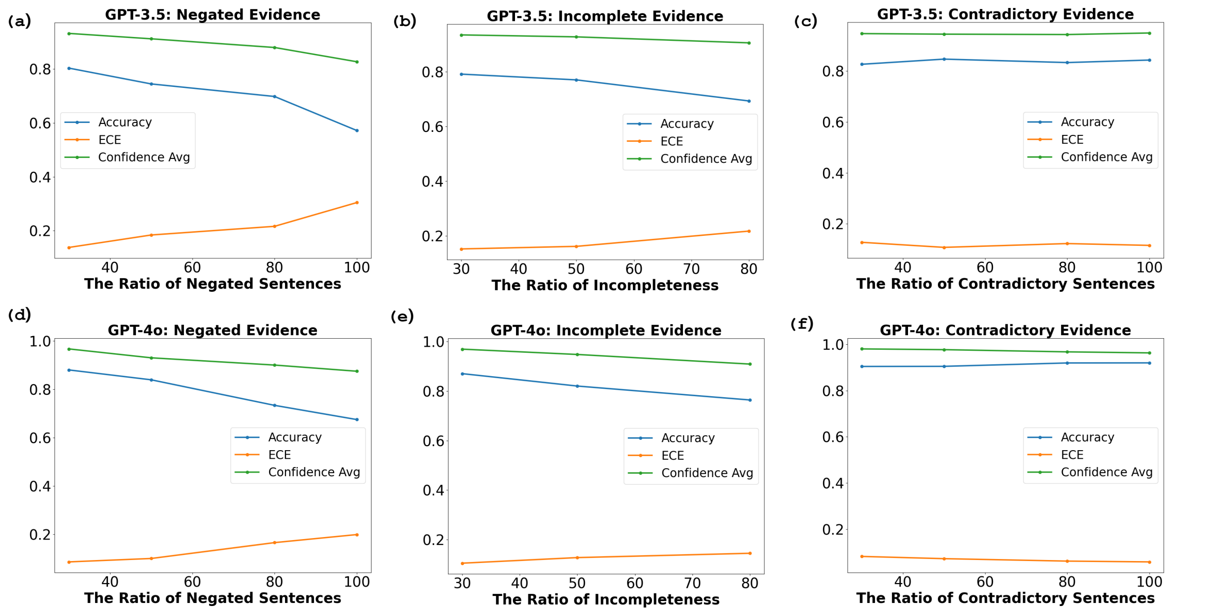

Figure 3: The results for the degree of variations in evidence for the SciQ dataset with verbalized method. We modified the number of negated sentences in negated evidence, sentences in incomplete evidence, and contradictory sentences in contradictory evidence (See Appendix C for entire results.).

an increase in the mean confidence for both models. This aligns well with the probability norm that stronger evidence increases the degree of belief. However, from the perspective of the calibration norm, these two types of evidence do not always positively contribute to accuracy or calibration. Rather, when low credible source and low detailed evidence were used, accuracy increased and ECE decreased. This suggests that in some cases, strong evidence may not be as useful as we expected for the language model to infer the correct answer. High confidence combined with low accuracy ultimately leads to overconfidence in incorrect predictions, resulting in high ECE.

Recent evidence, experimental evidence give confidence and accurate response in verbalized confidence On the other hand, as in (c), (d), (g), (h) in Figure 2, evidence containing the latest information or experiments showed higher confidence and accuracy compared to older information or observation-based evidence. Except for GPT-3.5 with experimental evidence, the ECE of stronger evidence was also lower, indicating that using stronger evidence in the cases of timeliness and experiments results in well-calibrated models. This means that in these cases, the language model utilizes the given evidence effectively without confusion, accurately reflecting the information in its predictions.

Through this experiment, we found that when stronger evidence is provided to the language model, it can significantly increase its verbalized confidence. However, this does not always lead to improvements in accuracy-related performance.

Token probability cannot reflect various degrees of reliability. Verbalized confidence is the only measure of confidence that, in all cases, appropriately and consistently increased in response to highly reliable evidence. As in Figure 4 in Appendix B), with token probability, confidence did not increase even when stronger evidence was presented. For example, with token probability, when the specificity of the evidence was altered or when the source's credibility was varied in GPT-4o, it failed to reflect confidence according to the strength of the evidence accurately. However, it accurately reflected reliability changes according to the source's credibility, timeliness, and whether an experiment was conducted in the evidence to its accuracy. Additionally, it showed a decrease in ECE in cases involving timeliness and experiments.

The sampling method can also generally reflect evidence. Although confidence decreased in the case of high specificity in GPT-4o, the sampling method overall showed higher confidence in high reliable evidence (See Figure 5 in Appendix B). Additionally, the sampling method showed higher accuracy of high reliable evidence in most cases except for specificity. We consider this phenomenon another positive aspect of self-consistent decoding (Wang et al., 2023b). A single response might not fully capture the reliability of evidence such as credibility, and timeliness, etc. However, multiple responses can increase the likelihood of accurately reflecting these aspects.

In the case of specificity, both verbalized confidence and sampling failed to reflect the concreteness of evidence in the responses properly. We interpreted that more detailed information can enhance confidence, but it also suggests that such excessive information may hinder the extraction of correct answers that match the question.

5 Ablation

LLMs tend to focus more on correct than incorrect information. In the no evidence and golden evidence cases, we interpreted that the language model possesses a certain degree of knowledge about the question in its parameters, and tends to be biased towards contexts aligned with this parametric knowledge rather than knowledge contradicting it, as seen win golden, contradictory and incomplete evidence. To justify this, we conducted an experiment adjusting the ratio of golden evidence in Figure 3. Figure 3 (a) and (d) show that as the number of original golden sentences decreases and the negated sentence increases, the performance of the language model gradually declines. However, it decreases significantly when there are no golden sentences left. Moreover, Figure 3 (b) and (e) demonstrate that as the original golden sentence decreases, performance decreases. On the other hand, Figure 3 (c) and (f) indicate that if the original sentence is sufficiently given, increasing the number of contradictory sentences does not affect the confidence and performance even if both of the contradictory evidences have the same sentence numbers. This shows that the language model focuses more on the given golden evidence in the context than inaccurate evidence, and this is why it maintains confidence and calibration despite incomplete and contradictory evidence.

Why do LLMs get confused by irrelevant context? Two interpretations are possible for the irrelevant case

1. The language model does not recognize irrelevant evidence which is different in content but the same in the field as irrelevant.

2. The language model considers irrelevant evidence as a kind of noise, which distracts the model and causes confusion.

To verify (1), instead of extracting irrelevant evidence from the same dataset, we used contexts from different datasets, for SciQ and TriviaQA dataset, we used evidence of GSM8K, and for GSM8K, using TriviaQA. As you can see in Figure 6 in Appendix D, even when using a new irrelevant, it did not completely match the completely irrelevant assumption. However, surprisingly, when using evidence from a completely different field, we found that the confidence, accuracy, and ECE metrics approached closer to the baseline no evidence case (P(H)) than when we used evidence where the content was different but the field was the same. This implies that the greater the irrelevance, the less the language model is distracted by the context. Therefore, we interpreted that there is a possibility that the language model satisfies the irrelevant assumption of Bayesian epistemology.

6 Conclusion

In this paper, we explored how changes in the informativeness and reliability of evidence affect the confidence and response of language models. Specifically, we examined how well language models stick to the probability and calibration norms outlined in Bayesian epistemology. We demonstrated that language models generally align well with Bayesian epistemology, especially when confidence is defined using verbalized confidence, which serves as an explicit introspection function in both confirmation tasks and strength of evidence tasks. This indicates that language models can be interpreted as possessing a belief in the view of Bayesian epistemology. At the same time, language models also exhibited a tendency to utilize information from unreasonable evidence, ignore inaccurate sentences, or let excessive information obstruct finding the right answers. Additionally, through ablation experiments of changing the ratio of golden evidence and negated sentences, we found that language models are more biased towards golden evidence, which can be seen as an epistemic characteristic of language models.

7 Limitations

In this paper, we investigated whether language models can distinguish and reflect various types of evidence in the inference stage. However, we did not focus on the deeper aspects, such as the training stage, architecture, and objective, which might have been the cause of the phenomenon in our findings. Why can language models ignore unreasonable contexts? Why do they focus more on generating answers based on correct information while disregarding the rest? Such deep analysis of the causes and future impacts of these character of language models are left for further research.

8 Ethics Statement

In the preparation of this paper, we utilized GPT-4o, for grammatical corrections and coding assistance. This technology served as an auxiliary resource to enhance the clarity and accuracy of our work, without directly influencing the research outcomes or decision-making processes involved. We acknowledge the support provided by OpenAI's GPT-4o in refining the presentation of our findings, ensuring that our use of this tool adheres to ethical guidelines and does not compromise the integrity of our research.

References

Robert Audi. 1997. *Epistemology: A Contemporary Introduction to the Theory of Knowledge*. Routledge, New York.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations.

Jiuhai Chen, Lichang Chen, Chen Zhu, and Tianyi Zhou. 2023. How many demonstrations do you need for in-context learning?

Yew Ken Chia, Guizhen Chen, Luu Anh Tuan, Soujanya Poria, and Lidong Bing. 2023. Contrastive chain-of-thought prompting.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram

Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya,

Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini: A family of highly capable multimodal models.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks.

Alan Hájek and Stephan Hartmann. 2010. Bayesian epistemology. In DancyJ, editor, *A Companion to Epistemology*. Blackwell.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2023. Methods for measuring, updating, and visualizing factual beliefs in language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2714–2731, Dubrovnik, Croatia. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Paul Horwich. 1982. *Probability and Evidence*. Cambridge University Press, Cambridge.

Colin Howson. 2000. *Hume's Problem: Induction and the Justification of Belief*. Oxford University Press, New York.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ArXiv*, abs/2311.05232.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know.

Nora Kassner, Oyvind Tafjord, Ashish Sabharwal, Kyle Richardson, Hinrich Schuetze, and Peter Clark. 2023. Language models with rationality.

Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. 2021. Beliefbank: Adding memory to a pre-trained language model for a systematic notion of belief.

Minsu Kim and James Thorne. 2024. Epistemology of language models: Do language models have holistic knowledge?

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Measuring faithfulness in chain-of-thought reasoning.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585, Singapore. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. 2024. Trustllm: Trustworthiness in large language models.

William Talbott. 2006. Bayesian epistemology. In Edward Zalta, editor, *Stanford Encyclopedia of Philosophy*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. 2023a. Fine-tuning language models for factuality.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023b. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting.

Bram M. A. van Dijk, Tom Kouwenhoven, Marco R. Spruit, and Max J. van Duijn. 2023. Large language models: The need for nuance in current debates and a pragmatic perspective on understanding.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. Towards understanding chain-of-thought prompting: An empirical study of what matters.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Jon Williamson. 2010. *In Defence of Objective Bayesianism*. Oxford University Press.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. *ArXiv*, abs/2103.15025.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Cheng Zhou, Xinbing Wang, and Luoyi Fu. 2023. Enhancing uncertainty-based hallucination detection with stronger focus. *ArXiv*, abs/2311.13230.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. 2023. Slic-hf: Sequence likelihood calibration with human feedback.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.

Appendix

A Results of Confirmation task

	Dataset	Metric	No_EVI	EVI	Coincidence	Irrelevant	Negation	Incomplete	Contradiction
GPT-3.5-turbo	SciQ	Confidence	0.671	0.781	0.785	0.594	0.638	0.723	0.764
		Accuracy ↑	0.676	0.829	0.839	0.526	0.6	0.792	0.837
		ECE ↓	0.312	0.171	0.154	0.44	0.381	0.205	0.16
	Trivia	Confidence	0.834	0.864	0.894	0.699	0.759	0.843	0.849
		Accuracy ↑	0.858	0.872	0.976	0.653	0.742	0.851	0.857
		ECE ↓	0.134	0.127	0.127	0.324	0.251	0.141	0.139
	GSM8K	Confidence	0.218	0.932	0.933	0.172	0.738	0.765	0.801
		Accuracy ↑	0.098	0.961	0.852	0.068	0.028	0.677	0.755
		ECE ↓	0.777	0.046	0.148	0.725	0.939	0.299	0.222
GPT-4o	SciQ	Confidence	0.621	0.799	0.833	0.565	0.653	0.744	0.813
		Accuracy ↑	0.711	0.92	0.905	0.675	0.655	0.835	0.925
		ECE ↓	0.276	0.082	0.1	0.314	0.334	0.165	0.078
	Trivia	Confidence	0.837	0.916	0.911	0.824	0.824	0.889	0.91
		Accuracy ↑	0.944	0.955	0.99	0.905	0.82	0.94	0.95
		ECE ↓	0.06	0.047	0.01	0.088	0.173	0.064	0.05
	GSM8K	Confidence	0.354	0.865	0.54	0.299	0.372	0.755	0.842
		Accuracy ↑	0.249	0.97	0.505	0.227	0.191	0.83	0.955
		ECE ↓	0.715	0.03	0.473	0.697	0.74	0.157	0.037

Table 2: The result of confirmation task with token probability method. We used 200 samples for GPT-4o due to the cost limit. NO_EVI refers the question with no context which means $P(H \mid \theta)$, serving as baseline. Others are the case of $P(H \mid E, \theta)$ where evidence appears in the context. EVI refers to the context in which the golden evidence from the dataset is given, while the other evidence types are those mentioned in section 3.2.

	Dataset	Metric	No_EVI	EVI	Coincidence	Irrelevant	Negation	Incomplete	Contradiction
GPT-3.5-turbo	SciQ	Confidence	0.874	0.921	0.916	0.798	0.828	0.888	0.922
		Accuracy ↑	0.693	0.846	0.853	0.551	0.617	0.777	0.853
		ECE ↓	0.18	0.076	0.077	0.248	0.211	0.111	0.074
	Trivia	Confidence	0.921	0.939	0.963	0.822	0.862	0.924	0.934
		Accuracy ↑	0.869	0.884	0.979	0.668	0.693	0.856	0.884
		ECE ↓	0.057	0.059	0.034	0.154	0.17	0.072	0.076
	GSM8K	Confidence	0.422	0.986	0.977	0.377	0.838	0.86	0.848
		Accuracy ↑	0.12	0.967	0.849	0.059	0.028	0.716	0.756
		ECE ↓	0.302	0.036	0.138	0.318	0.81	0.144	0.091
GPT-4o	SciQ	Confidence	0.872	0.968	0.959	0.852	0.871	0.923	0.965
		Accuracy ↑	0.694	0.934	0.924	0.708	0.698	0.84	0.933
		ECE ↓	0.18	0.06	0.102	0.149	0.114	0.132	0.066
	Trivia	Confidence	0.845	0.973	0.973	0.943	0.918	0.966	0.97
		Accuracy ↑	0.945	0.969	0.99	0.924	0.843	0.924	0.959
		ECE ↓	0.053	0.026	0.016	0.04	0.122	0.042	0.038
	GSM8K	Confidence	0.506	0.958	0.684	0.481	0.529	0.875	0.957
		Accuracy ↑	0.3	0.969	0.587	0.257	0.224	0.829	0.969
		ECE ↓	0.206	0.065	0.156	0.224	0.305	0.103	0.051

Table 3: The result of confirmation task with sampling method. We used 200 samples for GPT-4o due to the cost limit. NO_EVI refers the question with no context which means $P(H \mid \theta)$, serving as baseline. Others are the case of $P(H \mid E, \theta)$ where evidence appears in the context. EVI refers to the context in which the golden evidence from the dataset is given, while the other evidence types are those mentioned in section 3.2.

B Results of Strength of evidence task

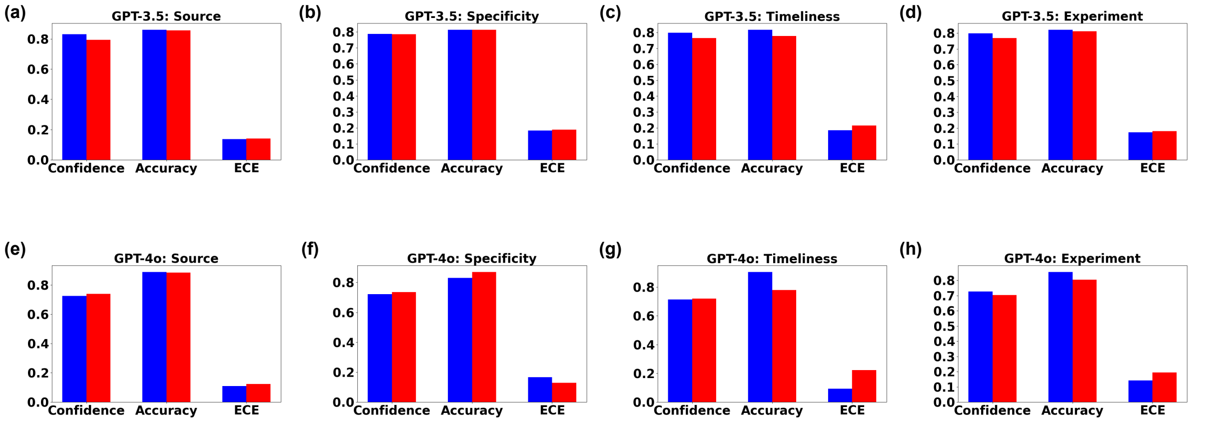

Figure 4: The results of the Strength of Evidence task on the SciQ dataset with token probability method. The blue bar represents the cases where the strength of evidence is high. Specifically, the blue bar indicates the context from more credible sources, more specific, recent, and experimental evidence, while the red color represents less credible sources, less specific, old, and observational evidence.

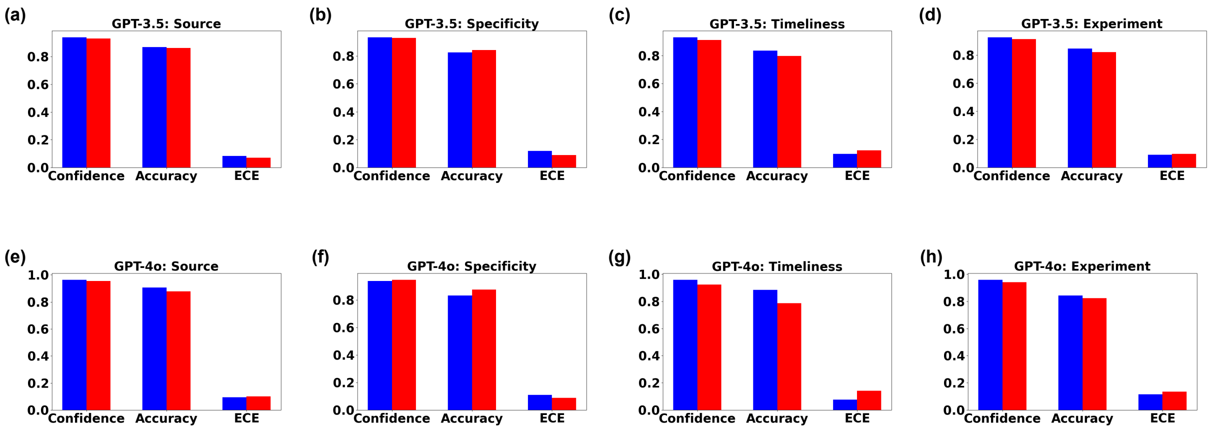

Figure 5: The results of the Strength of Evidence task on the SciQ dataset with sampling method. The blue bar represents the cases where the strength of evidence is high. Specifically, the blue bar indicates the context from more credible sources, more specific, recent, and experimental evidence, while the red color represents less credible sources, less specific, old, and observational evidence.

C Results of Ablation study on the ratio of golden evidence

	Dataset	Metric	Neg_30	Neg_50	Neg_80	Neg_100	Incomplete_30	Incomplete_50	Incomplete_80	Contradict_30	Contradict_50	Contradict_80	Contradict_100
GPT-3.5-turbo	SciQ	Confidence	0.932	0.912	0.88	0.827	0.935	0.928	0.906	0.947	0.945	0.943	0.95
		Accuracy ↑	0.803	0.744	0.745	0.572	0.791	0.77	0.693	0.827	0.847	0.833	0.843
		ECE ↓	0.138	0.184	0.216	0.304	0.152	0.161	0.216	0.127	0.108	0.122	0.115
	Trivia	Confidence	0.908	0.887	0.851	0.797	0.909	0.897	0.872	0.922	0.925	0.923	0.925
		Accuracy ↑	0.859	0.843	0.785	0.702	0.867	0.86	0.839	0.874	0.869	0.857	0.864
		ECE ↓	0.072	0.087	0.136	0.211	0.049	0.058	0.07	0.07	0.076	0.09	0.085
	GSM8K	Confidence	0.961	0.956	0.949	0.931	0.98	0.96	0.938	0.95	0.949	0.959	0.974
		Accuracy ↑	0.772	0.5	0.267	0.023	0.853	0.666	0.361	0.796	0.777	0.791	0.761
		ECE ↓	0.203	0.466	0.685	0.912	0.135	0.307	0.578	0.197	0.195	0.197	0.234
GPT-4o	SciQ	Confidence	0.967	0.93	0.9	0.875	0.969	0.948	0.909	0.98	0.977	0.968	0.963
		Accuracy ↑	0.88	0.839	0.734	0.675	0.87	0.82	0.764	0.904	0.905	0.92	0.92
		ECE ↓	0.087	0.101	0.166	0.2	0.105	0.128	0.145	0.082	0.072	0.062	0.058
	Trivia	Confidence	0.919	0.891	0.884	0.866	0.927	0.909	0.882	0.934	0.927	0.925	0.925
		Accuracy ↑	0.96	0.92	0.915	0.86	0.96	0.945	0.925	0.945	0.955	0.96	0.944
		ECE ↓	0.041	0.035	0.032	0.048	0.035	0.036	0.048	0.021	0.037	0.035	0.039
	GSM8K	Confidence	0.87	0.855	0.852	0.882	0.982	0.96	0.964	0.971	0.957	0.952	0.951
		Accuracy ↑	0.795	0.64	0.27	0.165	0.935	0.774	0.585	0.94	0.96	0.97	0.935
		ECE ↓	0.189	0.318	0.648	0.718	0.065	0.186	0.379	0.031	0.013	0.018	0.026

Table 4: The result of the ratio of golden sentence ablation study with verbalized method. We used 200 samples for GPT-4o due to the cost limit. We modified the number of negated sentences, the number of sentences in incomplete evidence, and the number of contradictory sentences in contradictory evidence and measured Confidence, Accuracy, and ECE. For example, Neg_80 means 80% of the entire sentences have been replaced into negated sentences, and Incomplete_80 means 80% of sentences have been deleted. Additionally, Contradict_80 refers 80% of evidence has been negated and appended to the evidence.

	Dataset	Metric	Neg_30	Neg_50	Neg_80	Neg_100	Incomplete_30	Incomplete_50	Incomplete_80	Contradict_30	Contradict_50	Contradict_80	Contradict_100
GPT-3.5-turbo	SciQ	Confidence	0.745	0.725	0.677	0.638	0.751	0.723	0.684	0.764	0.764	0.765	0.76
		Accuracy ↑	0.785	0.746	0.693	0.6	0.792	0.741	0.68	0.831	0.837	0.85	0.846
		ECE ↓	0.2	0.238	0.297	0.381	0.205	0.245	0.308	0.164	0.16	0.152	0.151
	Trivia	Confidence	0.854	0.831	0.8	0.759	0.853	0.843	0.822	0.878	0.849	0.851	0.857
		Accuracy ↑	0.863	0.808	0.742	0.668	0.851	0.852	0.814	0.873	0.857	0.867	0.851
		ECE ↓	0.2	0.187	0.251	0.326	0.146	0.141	0.178	0.132	0.139	0.136	0.147
	GSM8K	Confidence	0.877	0.807	0.765	0.738	0.894	0.765	0.532	0.842	0.801	0.796	0.801
		Accuracy ↑	0.803	0.518	0.262	0.028	0.881	0.677	0.384	0.825	0.775	0.777	0.741
		ECE ↓	0.207	0.469	0.725	0.939	0.118	0.299	0.534	0.172	0.222	0.211	0.257
GPT-4o	SciQ	Confidence	0.778	0.751	0.712	0.653	0.785	0.744	0.669	0.822	0.813	0.824	0.828
		Accuracy ↑	0.885	0.84	0.78	0.655	0.88	0.835	0.775	0.925	0.925	0.925	0.92
		ECE ↓	0.116	0.169	0.236	0.334	0.12	0.165	0.216	0.075	0.078	0.074	0.077
	Trivia	Confidence	0.905	0.85	0.853	0.824	0.911	0.889	0.858	0.913	0.91	0.914	0.918
		Accuracy ↑	0.94	0.9	0.86	0.82	0.95	0.94	0.925	0.96	0.95	0.945	0.944
		ECE ↓	0.058	0.104	0.146	0.173	0.045	0.064	0.078	0.04	0.05	0.055	0.052
	GSM8K	Confidence	0.765	0.611	0.421	0.372	0.856	0.755	0.599	0.856	0.842	0.851	0.862
		Accuracy ↑	0.835	0.61	0.351	0.191	0.945	0.83	0.59	0.95	0.955	0.965	0.955
		ECE ↓	0.16	0.349	0.614	0.74	0.055	0.127	0.393	0.05	0.037	0.035	0.041

Table 5: The result of the ratio of golden sentence ablation study with token probability. We used 200 samples for GPT-4o due to the cost limit. We modified the number of negated sentences, the number of sentences in incomplete evidence, and the number of contradictory sentences in contradictory evidence and measured Confidence, Accuracy, and ECE. For example, Neg_80 means 80% of the entire sentences have been replaced into negated sentences, and Incomplete_80 means 80% of sentences have been deleted. Additionally, Contradict_80 refers 80% of evidence has been negated and appended to the evidence.

	Dataset	Metric	Neg_30	Neg_50	Neg_80	Neg_100	Incomplete_30	Incomplete_50	Incomplete_80	Contradict_30	Contradict_50	Contradict_80	Contradict_100
GPT-3.5-turbo	SciQ	Confidence	0.904	0.885	0.865	0.828	0.906	0.888	0.87	0.914	0.922	0.921	0.918
		Accuracy ↑	0.822	0.77	0.706	0.616	0.813	0.777	0.71	0.859	0.853	0.856	0.852
		ECE ↓	0.091	0.115	0.158	0.211	0.093	0.111	0.165	0.064	0.074	0.072	0.07
	Trivia	Confidence	0.927	0.917	0.885	0.862	0.929	0.924	0.905	0.935	0.934	0.936	0.931
		Accuracy ↑	0.864	0.829	0.776	0.693	0.866	0.856	0.83	0.882	0.884	0.869	0.863
		ECE ↓	0.069	0.093	0.129	0.17	0.067	0.072	0.085	0.058	0.076	0.078	0.072
	GSM8K	Confidence	0.937	0.883	0.849	0.838	0.949	0.924	0.656	0.874	0.848	0.845	0.861
		Accuracy ↑	0.805	0.531	0.267	0.028	0.896	0.856	0.417	0.802	0.757	0.736	0.722
		ECE ↓	0.133	0.352	0.583	0.81	0.06	0.072	0.239	0.079	0.092	0.123	0.152
GPT-4o	SciQ	Confidence	0.943	0.922	0.904	0.871	0.954	0.923	0.906	0.958	0.965	0.959	0.957
		Accuracy ↑	0.893	0.848	0.807	0.698	0.887	0.84	0.77	0.929	0.933	0.938	0.934
		ECE ↓	0.078	0.109	0.114	0.187	0.086	0.132	0.137	0.063	0.066	0.045	0.075
	Trivia	Confidence	0.969	0.959	0.942	0.918	0.97	0.966	0.954	0.97	0.97	0.965	0.974
		Accuracy ↑	0.969	0.919	0.872	0.843	0.98	0.924	0.934	0.949	0.959	0.954	0.974
		ECE ↓	0.018	0.078	0.075	0.122	0.035	0.042	0.028	0.028	0.038	0.046	0.027
	GSM8K	Confidence	0.862	0.742	0.581	0.529	0.943	0.875	0.741	0.948	0.957	0.952	0.944
		Accuracy ↑	0.882	0.685	0.407	0.224	0.943	0.829	0.622	0.964	0.969	0.964	0.954
		ECE ↓	0.059	0.074	0.174	0.305	0.047	0.103	0.119	0.067	0.051	0.063	0.08

Table 6: The result of the ratio of golden sentence ablation study with sampling method. We used 200 samples for GPT-4o due to the cost limit. We modified the number of negated sentences, the number of sentences in incomplete evidence, and the number of contradictory sentences in contradictory evidence. For example, Neg_80 means 80% of the entire sentences have been replaced into negated sentences, and Incomplete_80 means 80% of sentences have been deleted. Additionally, Contradict_80 refers 80% of evidence has been negated and appended to the evidence.

D Results of Ablation study on irrelevant evidence

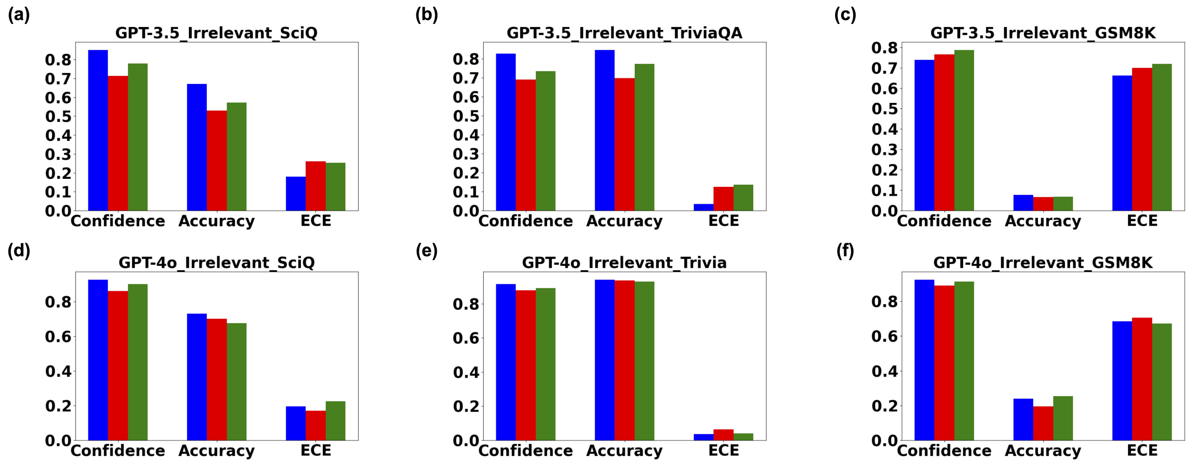

Figure 6: The results of ablation for irrelevant evidence. The blue bar represents the result of no evidence $P(H)$, serving as a baseline. The red bar results from irrelevant evidence by replacing evidence from other samples within the same dataset explained in section 3.2. The green bar represents irrelevant evidence from another dataset.

E Experimental Detail

E.1 Hyperparameter

We utilized OpenAI's API to create a dataset containing evidence and conducted inference experiments. Specifically, we used GPT-4-0613 to generate Negated evidence, Coincidental evidence, and Contradictory evidence required for the confirmation task, and gpt-4o-2024-05-13 to create evidence necessary for the strength of evidence. The inference was performed using GPT-3.5-0125 and GPT-4o-2024-05-13 with settings of temperature=1.0 and top_p=1.0.

E.2 Evaluation Detail

According to (Kuhn et al., 2023), for the SciQ and TriviaQA datasets, we considered a model's response as correct if its Rouge-L score (Lin, 2004) with the golden label is 0.3 or higher. For GSM8K, only responses that were an exact match with the golden label were considered correct.

For sampling method for measuring confidence, we set the ratio of most frequent response as the confidence. As the datasets are open-ended question, we should consider the synonym of each responses. In order to handle this, we used GPT-4o-2024-05-13 to capture the semantic similarity and calculate the frequency of the most common response.

E.3 Dataset

For SciQ and GSM8K, we extracted the samples containing the explanation, including more than 4 sentences to create various proportions of negated sentences in the ablation study. Similarly, for trivia QA, we used the explanation[1] including more than 4 sentences and extracting 1000 samples. We generated negated sentences using GPT-4-0613 for negated and contradictory evidence and filtered out samples containing incorrect sentences. Similarly, we used GPT-4o-2024-05-13 for generating Strength of Evidence task and also filtered out the generated strength of evidence that included a wrong template. The total number of samples is shown in Table 7 and Table 8. We used all these samples when inferencing with GPT-3.5-turbo and 200 samples for GPT-4o-2024-05-13.

[1]We used the context of each question as evidence. For the context of each sample, we used the positive passage in https://huggingface.co/datasets/Tevatron/wikipedia-trivia.

	NO_EVI	EVI	Coincidence	Irrelevant	Negation	Incomplete	Contradiction
SciQ	1095	1095	1095	1095	991	1095	991
TriviaQA	1000	1000	1000	1000	798	1000	798
GSM8K	622	622	622	622	618	622	618

Table 7: The number of samples for the Confirmation task dataset.

	High Credible Source	Low Credible Source	High Specificity	Low Specificity	Recent	Old	Experiment	Observation
SciQ	1095	1095	1093	1093	1074	1074	1094	1094

Table 8: The number of samples for the Strength of evidence task dataset.

F Prompt

In this section, we will show the prompt for inference,

F.1 Prompt for Inference

Verbal Confidence Prompt

Provide your best guess and the probability that it is correct (0.0 to 1.0) for the following question based on the evidence.
Give ONLY the guess and probability, no other words or explanation.
For example
Guess: <most likely guess, as short as possible; not a complete sentence, just the guess!>
Probability: <the probability between 0.0 and 1.0 that your guess is correct based on the given evidence , without any extra commentary whatsoever; just the probability!>
###**The question**: {question}
###**The evidence**: {evidence}

Table 9: A prompt for verbal confidence and guess of answer from language models. We follow (Tian et al., 2023b).

Prompt for Token probability and Sampling

Provide your best guess for the following question based on the evidence.
Give ONLY the guess, no other words or explanation.
For example
Guess: <most likely guess, as short as possible; not a complete sentence, just the guess!>
###**The question**: {question}
###**The evidence**: {evidence}

Table 10: A prompt for Token probability and guess of answer from language models. We do not need to extract the confidence by prompt, so all we need is to extract the guess.

F.2 Prompt for Generating Evidence

Prompt for Negating the evidence

###**Example**: "Biochemical reactions of metabolism can be divided into two general categories: catabolic reactions and anabolic reactions. You can watch an animation showing how the two categories of reactions are related at this URL: http://classes. midlandstech. edu/carterp/courses/bio225/chap05/lecture1. htm."
Revise or negate each sentence in the ###Example with incorrect information yet relevant information. The response ###Negation should have same number of sentence with ###Example.
###**Negation**: "Biochemical reactions of metabolism are typically classified into only one category: equilibrium reactions. You can view a static image illustrating the isolated function of equilibrium reactions at this URL: http://classes.midlandstech.edu/carterp/courses/bio225/chap05/lecture2.htm."
###**Example**: "An anaerobic organism is any organism that does not need oxygen for growth and even dies in its presence. Obligate anaerobes will die when exposed to atmospheric levels of oxygen. Clostridium perfringens bacteria, which are commonly found in soil around the world, are obligate anaerobes. Infection of a wound by C. perfringens bacteria causes the disease gas gangrene. Obligate anaerobes use molecules other than oxygen as terminal electron acceptors."
Revise or negate each sentence in the ###Example with incorrect information yet relevant information. The response ###Negation should have same number of sentence with ###Example.
###**Negation**: "An anaerobic organism is any organism that requires oxygen for growth and thrives in its presence. Obligate aerobes will perish when deprived of atmospheric oxygen levels. Staphylococcus aureus bacteria, which are rarely found in aquatic environments, are obligate aerobes. Infection of a wound by S. aureus bacteria causes the disease known as athlete's foot. Obligate aerobes use molecules such as hydrogen or sulfur as terminal electron acceptors."
###**Example**: "The energy of a mechanical wave can travel only through matter. The matter through which the wave travels is called the medium (plural , media). The medium in the water wave pictured above is water, a liquid. But the medium of a mechanical wave can be any state of matter, even a solid."
Revise or negate each sentence in the ###Example with incorrect information yet relevant information. The response ###Negation should have same number of sentence with ###Example.
###**Negation**: "The energy of a mechanical wave can travel through both matter and vacuum. The space through which the wave travels is termed the conduit. The conduit in the water wave pictured above is air, a gas. However, the conduit of a mechanical wave can be exclusively in a gaseous state, not a solid or liquid."
###**Example**: "What group of animals begins its life in the water, but then spends most of its life on land? Amphibians! Amphibians are a group of vertebrates that has adapted to live in both water and on land. Amphibian larvae are born and live in water, and they breathe using gills. The adults live on land for part of the time and breathe both through their skin and with their lungs as their lungs are not sufficient to provide the necessary amount of oxygen."
Revise or negate each sentence in the ###Example with incorrect information yet relevant information. The response ###Negation should have same number of sentence with ###Example.
###**Negation**: "What group of animals begins its life on land, but then spends most of its life in water? Reptiles! Reptiles are a group of vertebrates that has adapted to live mainly on land but also in water. Reptile eggs are laid and hatch on land, and they breathe using lungs from birth. The adults live in water for part of the time and breathe exclusively through their lungs as their skin is not permeable enough to facilitate breathing."
###**Example**:{source}
Revise or negate each sentence in the ###Example with incorrect information yet relevant information. The response ###Negation should have same number of sentence with ###Example.
###**Negation**:

Table 11: A prompt for negating the each sentence in golden evidence.

Prompt for Token probability and Sampling

###**Question**: "What does the pull of the moon's gravity on earth cause?", Answer: tides
Make explanation that realize the answer to the question by knowing from unreasonable guess. You should not express that answer from unreasonable guess is correct for the question.
###**Explanation**: "Interestingly, while reviewing some casual photographs taken during a lunar eclipse, a curious pattern emerged that reminded me of a physics class discussion. The photos captured how remarkably calm and aligned everything seemed under the moonlight, which brought to mind the effects of the moon's gravity on Earth. It dawned on me, almost by happenstance, that the moon's gravitational pull is what primarily drives the ocean tides. This realization was unexpected, as the focus of the eclipse viewing was purely for its beauty and not the scientific implications. It's fascinating how a moment of leisure led to a reminder that while the moon's gravity tugs at the land, only the water, being fluid, visibly responds, creating the tides we observe."

###**Question**: "What living species of bird is considered the largest of the raptors?" Answer: golden eagles Make explanation that realize the answer to the question by knowing from unreasonable guess. You should not express that answer from unreasonable guess is correct for the question.
###**Explanation**: "During a casual discussion at a local café in Northern California, after a day of hiking, I was sharing stories and showing photos I had taken of a large bird, Golden Eagle, I encountered. Not being an expert, I half-jokingly guessed it might be one of the larger bird species due to its impressive size and wingspan."

###**Question**: "What is another term for blood clotting?" Answer: coagulation
Make explanation that realize the answer to the question by knowing from unreasonable guess. You should not express that answer from unreasonable guess is correct for the question.
###**Explanation**:"At a community health fair, while casually discussing various health topics, an attendee mentioned his surprise at how quickly a minor cut he had stopped bleeding. Not being a healthcare professional, I playfully suggested it might be due to his body being exceptionally good at 'coagulating'—a term I recalled from a movie rather than a biology class."

###**Question**: "What is the protein in red blood cells which transports oxygen around the body?", Answer: hemoglobin
Make explanation that realize the answer to the question by knowing from unreasonable guess. You should not express that answer from unreasonable guess is correct for the question.
###**Explanation**: "During a spirited trivia night with friends, where various odd facts were thrown around, someone challenged the group to name the protein in red blood cells responsible for transporting oxygen. Without any medical background, I haphazardly threw out 'hemoglobin' as a guess, mainly because it sounded scientific and related to blood. My guess was influenced by overhearing a conversation earlier that day at the gym where someone mentioned they were taking iron supplements to improve their 'hemoglobin levels' for better endurance"

###**Question**: {question}, **Answer**: {answer}
Make explanation that realize the answer to the question by knowing from unreasonable guess. You should not express that answer from unreasonable guess is correct for the question.
###**Explanation**:

Table 12: A prompt for making coincidental evidence.

Prompt for Making credible evidence

###**question**: "What substance does the phillosopher stone change the base material to?"
###**answer**: "gold"

For this ###question, ###answer pairs, make 3 evidences with difference power of evidence in the aspect of Source Credibility.

###**Highly Credible Source**: "A leading professor of alchemy at a renowned university published a peer-reviewed paper documenting the transmutation of lead into gold using the Philosopher's Stone."

###**Moderately Credible Source**: "A respected independent alchemist reported successful transmutations in his personal journal."

###**Low Credibility Source**: "An anonymous blog post claims to have discovered the Philosopher's Stone and successfully converted lead into gold."

###**question**: "Compounds with aluminum and silicon are commonly found in the clay fractions of soils derived from what?"
###**answer**: "volcanic ash"

For this ###question, ###answer pairs, make 3 evidences with difference power of evidence in the aspect of Source Credibility.

###**Highly Credible Source**:"A peer-reviewed study published in the Journal of Soil Science by researchers from a top-tier university provides detailed analysis and evidence that clay fractions in soils derived from volcanic ash predominantly contain compounds of aluminum and silicon."

###**Moderately Credible Source**:"A detailed report by a well-known geologist in a respected geology magazine discusses the mineral composition of clay fractions in soils and highlights volcanic ash as a common origin of aluminum and silicon compounds."

###**Low Credibility Source**:"A gardening enthusiast's blog post mentions that soils rich in aluminum and silicon compounds often come from volcanic ash, based on their personal observations and informal tests."

###**question**: {question}
###**answer**: {answer}

For this ###question, ###answer pairs, make 3 evidences with difference power of evidence in the aspect of Source Credibility.

Table 13: The prompt for generating various of evidence according to credibility. We did not use moderate credibility evidence, as it is similar to other evidence.

Prompt for Making specificity evidence

###**question**: "What substance does the phillosopher stone change the base material to?"
###**answer**: "gold"

For this ###question, ###answer pairs, make 3 evidences with difference power of evidence in the aspect of Specificity and detail.

###**Highly Specific Evidence**: "Detailed records from 16th-century experiments show precise measurements and procedures for transmuting lead into gold using a substance identified as the Philosopher's Stone."

###**Moderately Specific Evidence**: "Historical documents suggest that some alchemists reported converting metals into gold, but the details are sparse."

###**General Evidence**: "There are general mentions in ancient texts about the ability to convert base metals into gold."

###**question**: "Compounds with aluminum and silicon are commonly found in the clay fractions of soils derived from what?"
###**answer**: "volcanic ash"

For this ###question, ###answer pairs, make 3 evidences with difference power of evidence in the aspect of Specificity and detail.

###**Highly Specific Evidence**:"Geochemical analyses of soil samples from regions with known volcanic activity demonstrate that the clay fractions are predominantly composed of alumino-silicate minerals, confirming that these soils are derived from volcanic ash deposits."

###**Moderately Specific Evidence**:"Scientific studies indicate that soils in volcanic regions frequently contain clay fractions rich in aluminum and silicon compounds, which suggests a derivation from volcanic ash."

###**General Evidence**:"Many references in soil science literature mention that clay fractions with aluminum and silicon are often associated with volcanic ash origins."

###**question**: {question}
###**answer**: {answer}

For this ###question, ###answer pairs, make 3 evidences with difference power of evidence in the aspect of Specificity and detail.

Table 14: The prompt for generating various evidence according to specificity. We did not use moderate specific evidence, as it is similar to other evidence

Prompt for Making timeliness evidence

###**question**: "What substance does the phillosopher stone change the base material to?"
###**answer**: "gold"

For this ###question, ###answer pairs, make 2 evidences with difference power of evidence in the aspect of timeliness. (the older evidence should be before 18th-century)

###**Recent Evidence**: "A 2022 study published in a scientific journal provides new experimental data supporting the possibility of metal transmutation using a newly synthesized substance resembling the Philosopher's Stone."

###**Older Evidence**: "A 17th-century manuscript claims to have witnessed the transformation of base metals into gold using an alchemical process."

###**question**: "Compounds with aluminum and silicon are commonly found in the clay fractions of soils derived from what?"
###**answer**: "volcanic ash"

For this ###question, ###answer pairs, make 2 evidences with difference power of evidence in the aspect of timeliness. (the older evidence should be before 18th-century)

###**Recent Evidence**: "A 2019 study published in a geochemistry journal confirms that soils derived from volcanic ash predominantly contain clay fractions with high concentrations of aluminum and silicon compounds."

###**Older Evidence**: "A 16th-century agricultural text describes soils from regions with volcanic activity as rich in aluminosilicate clays, derived from the weathering of volcanic ash."

###**question**: {question}
###**answer**: {answer}

For this ###question, ###answer pairs, make 2 evidences with difference power of evidence in the aspect of timeliness. (the older evidence should be before 18th-century)

Table 15: The prompt for generating various evidence according to timeliness.

Prompt for Making experimental evidence

###**question**: "What substance does the phillosopher stone change the base material to?"
###**answer**: "gold"
For this ###question, ###answer pairs, make 2 evidences with different levels of strength in the aspect of Experimental or Observational Evidence, ensuring that the observational evidence includes direct observations from normal people such as "several witnesses observed."

###**Experimental Evidence**: "Recent laboratory experiments conducted under controlled conditions have demonstrated the conversion of lead into gold using a synthetic version of the Philosopher's Stone."

###**Observational Evidence**: "Several eyewitness accounts from the 1600s describe seeing alchemists successfully convert metals into gold, though these were not scientifically verified."

###**question**: "Compounds with aluminum and silicon are commonly found in the clay fractions of soils derived from what?"
###**answer**: "volcanic ash"

For this ###question, ###answer pairs, make 2 evidences with different levels of strength in the aspect of Experimental or Observational Evidence, ensuring that the observational evidence includes direct observations from normal people such as "several witnesses observed."

###**Experimental Evidence**: "A series of controlled soil analysis experiments have shown that soils formed from volcanic ash consistently contain high concentrations of aluminum and silicon compounds in their clay fractions."

###**Observational Evidence**: "Several teams have directly observed that soils in regions with volcanic activity, particularly those rich in clay, contain significant amounts of aluminum and silicon."
###**question**: {question}
###**answer**: {answer}

For this ###question, ###answer pairs, make 2 evidences with different levels of strength in the aspect of Experimental or Observational Evidence, ensuring that the observational evidence includes direct observations from normal people such as "several witnesses observed."

Table 16: The prompt for generating various evidence according to the existence of the experiment.